A Bayesian approach to Mendelian Randomisation with dependent instruments

Chin Yang Shapland[1,*], John R. Thompson[1], and Nuala A. Sheehan[1]

[1] *Department of Health Sciences and Genetics, University of Leicester, LE1 6TP, UK.*

[*] *Corresponding author: Oakfield House, Oakfield Grove, Clifton, BS8 2BN, UK. , E-mail: ew18103@bristol.ac.uk*

Revised version: 15 May 2018

Abstract

Mendelian randomisation (MR) is a method for establishing causality between a risk factor and an outcome by using genetic variants as instrumental variables. In practice the association between individual genetic variants and the risk factor is often weak, which may lead to a lack of precision in the MR and even biased MR estimates. Usually, the most significant variant within a genetic region is selected to represent the association with the risk factor, but there is no guarantee that this variant will be causal or that it will capture all of the genetic association within the region. It may be advantageous to use extra variants selected from the same region in the MR. The problem is to decide which variants to select. Rather than select a specific set of variants, we investigate the use of Bayesian model averaging (BMA) to average the MR over all possible combinations of genetic variants. Our simulations demonstrate that the BMA version of MR out-performs classical estimation with many dependent variants and performs much better than a MR based on variants selected by penalised regression. In further simulations we investigate robustness to violations in the model assumptions and demonstrate a sensitivity to the inclusion of invalid instruments. The method is illustrated by applying it to a MR of the effect of body mass index on blood pressure using SNPs in the *FTO* gene.

Keywords: Bayesian model averaging; Mendelian randomisation; many weak instruments; dependent SNPs

1 Introduction

In epidemiological studies, ordinary least squares (OLS) regression is often used to estimate the effect of a modifiable risk factor on an outcome of interest, but such an analysis is biased in the presence of unmeasured confounding. Mendelian randomisation (MR) adjusts for unmeasured confounding using genetic variants as instrumental variables. Each genetic instrument must (1) influence the risk factor, (2) be independent of the confounders and (3) associate with the outcome only through its effect on the risk factor [18, 33, 45, 60, 65]. The application of MR comes with its own challenges [19, 20, 66], one of which is that most genetic instruments only explain a small proportion of the variation in the risk factor; this is commonly known as the weak instruments problem. When the instruments are weak, a very large sample size will be required to provide power and precision [9, 56] and as a result it can be advantageous to include as many genetic variants as possible in the MR, provided that they are all valid instruments.

Usually the variants included in a MR are chosen to be independent of one another [11, 70], but there are situations in which it may be beneficial to include sets of variants from the same genetic region even though they will be in linkage disequilibrium (LD) and thus not independent. For instance, there may be several causal variants in the region, or it might be difficult to identify the causal variant, or the causal variant might not have been measured. In each case it could be beneficial to use a set of proxy variants.

Several Mendelian randomisation studies that have used SNPs from the same genetic region as instruments, these include, a study of plasma level on cardiovascular disease that used SNPs from the AHSG gene [26], a study of adiponectin and type 2 diabetes with SNPs from ADIPOQ [75], and a study of adiposity and cardiovascular disease with SNPs from TRIB1 [17]. However, these studies gave individual causal effect estimates from each SNP. Other Mendelian randomisation studies have combined dependent SNPs into a single instrument, using allele scores. However, due to a lack of external data on the genetic-exposure association, they either derived the weights for the allele score from the dataset under analysis or did not weight the risk allele. The former approach is subject to bias [12] and the latter to a loss in precision [22].

Bayesian approaches offer a systematic and structured way of incorporating external biological knowledge into the statistical analysis [1]. Many publications have discussed the biological justification for the genetic instruments used in the MR [32], so why not incorporate this information into the statistical analysis? In the context of Mendelian randomisation, the selection of instruments using the p-value or F-statistic cannot distinguish between situations in which (1) there is not sufficient data to detect an effect and (2) there is no effect. By computing the posterior effect probability for each variant, the Bayesian approach is able to quantify these two possibilities. Instead of selecting instruments based on significance, instruments that are biologically relevant to the exposure or instruments that have shown association with it in a meta-analysis, can be given more weight.

The aim of this paper is to investigate the use of Bayesian model averaging (BMA) [35] in the context of a MR with many dependent instruments. BMA allows the MR to consider all possible combinations of the genetic variants and combines the resulting causal effect estimates with appropriate weights, all within the same dataset. We perform simulations to identify factors that affect convergence and mixing, robustness and the performance of BMA compared with two-stage least squares (2SLS), limited information maximum likelihood (LIML) and a recently available penalised regression based method called, Some Invalid Some Valid Instrumental Variables Estimator (sisVIVE) [40]. The simulations will mimic the SNP patterns similar to the ones seen in the regional plots from the Schizophrenia Psychiatric GWAS Consortium [58], in which genes have a lead SNP (the most significant) and many SNPs correlated with it. Our robustness section will follow the procedure seen in O'Malley *et al.* [53], where they tested the sensitivity of the estimators to the instrumental variable and distributional assumptions. Finally, the paper will consider the use of real data from the GRAPHIC study [69] to estimate the causal effect of BMI on blood pressure using SNPs from the *FTO* gene as the instruments.

In this paper, models are fitted using the R package *ivbma*. We use lower case italics for the name of the package and upper case (IVBMA) for the general method of applying BMA to instrumental variable analyses such as Mendelian randomisation. Similarly we refer to the R package *sisVIVE* that implements the sisVIVE method.

2 Instrumental Variable Bayesian Model Averaging

Focusing a statistical analysis on a single pre-selected model has been described as a quiet scandal [7]. Bayesian model averaging (BMA) avoids the scandal by considering a range of models. Suppose that we decide to consider K possible models $M_1, \ldots, M_K$, then the posterior distribution of a quantity of interest Λ given data D is;

$$p(\Lambda|D) = \sum_{k=1}^{K} p(\Lambda|M_k, D)p(M_k|D). \tag{1}$$

This is an average of the posterior distributions under each of the models (M_k), weighted by their posterior model probability. In the context of a MR with many instruments, the models M_k represent different combinations of the potential set of genetic instruments. The posterior probability of model M_k is given by,

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{l=1}^{K} p(D|M_l)p(M_l)}. \tag{2}$$

where

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \tag{3}$$

is the integrated likelihood of model M_k, θ_k is the vector of parameters of model M_k, $p(\theta_k|M_k)$ is the prior density of θ_k under model M_k, $p(D|\theta_k, M_k)$ is the likelihood and $p(M_k)$ is the prior probability that M_k is the true model.

The posterior mean and variance of Λ are;

$$E[\Lambda|D] = \sum_{k=0}^{K} \hat{\Lambda}_k p(M_k|D),$$

$$\mathrm{Var}[\Lambda|D] = \sum_{k=0}^{K} (\mathrm{Var}[\Lambda|D, M_k] + \hat{\Lambda}_k^2)p(M_k|D) - E[\Lambda|D]^2,$$

where $\hat{\Lambda}_k = E[\Lambda|D, M_k]$.

Hoeting *et al.* [35] gives a comprehensive tutorial on BMA. Bayesian model averaging approaches have been adapted by econometricians for large numbers of exogenous variables, as a way to avoid over-fitting the regression model. Koop *et al.* [44] introduced BMA into the framework of instrumental variable analysis. They argued that investigators may be uncertain about whether their variables belong to the groups of endogenous variables, exogenous variables or instruments, and BMA would be a way of incorporating this uncertainty. Lenkoski *et al.* [46] have shown through simulations that, unlike its classical counterparts, instrumental variable Bayesian model averaging (IVBMA) does not suffer from many instrument bias.

IVBMA reduces weak instrument bias by averaging the estimated causal effect from models with different sets of instruments. The selection of instruments is conditional on the likelihood of the data and the given priors. IVBMA also gives the posterior probability of inclusion for each instrument and a posterior probability to measure support for the null hypothesis of no causal effect. As we rarely know the true causal variant, IVBMA offers a way of comparing multiple plausible models with different instruments without selection by p-value.

The main advantage of applying BMA to Mendelian randomisation is its potential to reduce the many weak instrument bias by allowing flexibility in instrument inclusion without introducing selection bias (from using the same dataset for calculating the weights of the instruments and estimating the causal effect). Karl *et al.* [41] designed an algorithm for instrumental variable Bayesian model averaging (IVBMA) and later wrote an R package, *ivbma*, to implement their approach. *ivbma* uses Markov Chain Monte Carlo Model Composition (MC3) within a Gibbs sampler, which is a special case of a Metropolis-within-Gibbs algorithm. MC3 can be considered as a Metropolis-Hastings step in the space of the models; MC3 moves through model space, accepting or rejecting a model via a Conditional Bayes Factor. The MC3-within-Gibbs sampler is particularly efficient when there are many potential models [51]. The *ivbma* model is;

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\tau} + \boldsymbol{\eta} \tag{4}$$
$$\boldsymbol{Y} = \boldsymbol{X}\beta_{XY} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{5}$$

where the outcome, $\boldsymbol{Y}$, and the endogenous risk factor, $\boldsymbol{X}$, are both $n \times 1$. $\boldsymbol{W}$ denotes an $n \times p$ matrix of further explanatory variables and $\boldsymbol{Z}$ contains the instrumental variables in an $n \times k$ matrix. $\binom{\varepsilon_i}{\eta_i} \sim N_2(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$

The MC3-within-Gibbs sampler as implemented in *ivbma* is described in full in Karl *et al.* [41] but briefly the algorithm creates a sequence for the full set of parameters $\theta^{(1)} \dots \theta^{(S)}$ where;

$$\theta^{(s)} = \left\{ \boldsymbol{\rho}^{(s)}, \mathcal{M}_{sec}^{(s)}, \boldsymbol{\lambda}^{(s)}, \mathcal{M}_{fst}^{(s)}, \Sigma^{(s)} \right\},$$

$\boldsymbol{\rho}^{(s)} = [\beta_{XY}, \boldsymbol{\gamma}]$ and $\boldsymbol{\lambda}^{(s)} = [\boldsymbol{\delta}, \boldsymbol{\tau}]$. $\mathcal{M}_{sec}$ and $\mathcal{M}_{fst}$ are the model spaces for Equation 5 and 4 respectively, i.e. the model space for $\mathcal{M}_{fst}$ includes Z and W, whereas $\mathcal{M}_{sec}$ includes X and W. Given the current state $\rho^{(s)}$ and data $\mathcal{D}$, the *ivbma* algorithm starts;

1. Sample $\mathcal{M}'_{sec}$ from the neighbourhood of $\mathcal{M}_{sec}^{(s)}$, i.e. models that differ from $\mathcal{M}_{sec}^{(s)}$ by one variable. Then calculate

$$\alpha = \frac{p(\mathcal{D}|\mathcal{M}'_{sec}, \boldsymbol{\lambda}^{(s)}, \Sigma^{(s)})}{p(\mathcal{D}|\mathcal{M}_{sec}^{(s)}, \boldsymbol{\lambda}^{(s)}, \Sigma^{(s)})} 1\left\{ \mathcal{M}'_{sec}, \mathcal{M}_{fst}^{(s)} \in \mathcal{A} \right\}$$

   with probability $min\{\alpha, 1\}$ set $\mathcal{M}_{sec}^{(s+1)} = \mathcal{M}'_{sec}$, otherwise $\mathcal{M}_{sec}^{(s+1)} = \mathcal{M}_{sec}^{(s)}$

2. Sample $\boldsymbol{\rho}^{(s+1)}$ from the conditional posterior distribution of $\boldsymbol{\rho}_{\mathcal{M}_{sec}^{(s+1)}}$, i.e. posterior distribution for coefficients of the new model in the second stage regression.

3. Sample $\mathcal{M}'_{fst}$ from the neighbourhood of $\mathcal{M}_{fst}^{(s)}$. Then calculate

$$\alpha = \frac{p(\mathcal{D}|\mathcal{M}'_{fst}, \boldsymbol{\rho}^{(s+1)}, \Sigma^{(s)})}{p(\mathcal{D}|\mathcal{M}_{fst}^{(s)}, \boldsymbol{\rho}^{(s+1)}, \Sigma^{(s)})} 1\left\{ \mathcal{M}_{sec}^{(s+1)}, \mathcal{M}'_{fst} \in \mathcal{A} \right\}$$

   with probability $min\{\alpha, 1\}$ set $\mathcal{M}_{fst}^{(s+1)} = \mathcal{M}'_{fst}$, otherwise $\mathcal{M}_{fst}^{(s+1)} = \mathcal{M}_{fst}^{(s)}$

4. Sample $\boldsymbol{\lambda}^{(s+1)}$ from the conditional posterior distribution of $\boldsymbol{\lambda}_{\mathcal{M}_{fst}^{(s+1)}}$, i.e. posterior distribution for coefficients of the new model in the first stage regression.

5. Use $\boldsymbol{\lambda}^{(s+1)}$ and $\boldsymbol{\rho}^{(s+1)}$ to calculate $\boldsymbol{\varepsilon}^{s+1}$ and $\boldsymbol{\eta}^{(s+1)}$ and sample $\boldsymbol{\Sigma}^{(s+1)}$ from the conditional posterior distribution of $\boldsymbol{\Sigma}$.

See Karl *et al.* [41] for the derivation of Bayes Factor in Steps (1) and (3) and the full equation of the conditional posterior distributions for each parameter.

The R package, *ivbma*, imposes the following priors on the parameters:

$$\begin{aligned}
[\beta_{XY}, \boldsymbol{\gamma}] &\sim N(0,1), \\
[\boldsymbol{\delta}, \boldsymbol{\tau}] &\sim N(0,1), \\
[\mathcal{M}_{fst}, \mathcal{M}_{sec}] &\sim Bern(0.5), \\
\boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, 3\right).
\end{aligned}$$

where N, $Bern$ and $\mathcal{W}^{-1}$ stands for the Normal, Bernoulli and Inverse-Wishart distribution respectively. Note that giving $Bern(0.5)$ distribution to the model spaces tells *ivbma* that each possible model, is equally likely to be true, where each model contains different covariates. *ivbma* does not allow these priors to be altered. Lenkoski *et al.* [46] provide a description of the options available in the R package.

3 Method of Simulation

The human genome is made up of 3.2×10^9 base-pairs(bp) and the average length of protein-coding genes is 53.6×10^3 [63]. The International HapMap Project estimated 10 million SNPs within the human genome [64] and 5 million of these have allele frequency greater than 10% [57]. Based on these figures we estimated that a typical gene might contain 80 or 90 common SNPs.

The simulation study used two methods. The first was designed to investigate the effect of different LD patterns and minor allele frequency (MAF) using controlled but artificial patterns of linkage disequilibrium (LD). The second method used the GENOME [47] program to create realistic but uncontrolled patterns of LD and MAF. In all scenarios genotype, exposure and outcome of interest were simulated for a study of 2,000 individuals.

3.1 Artificial LD

We consider four artifical patterns of LD as illustrated in Figure 1. Patterns I, II and III assume one functional variant located in the middle of the region. Pattern IV divides the region into two and places one causal SNP at the centre of each part. Suppose that k is the total number of SNPs or potential instruments, n is the number of individuals, $\boldsymbol{\rho_{max}}$ is the maximum correlation between the genotype of the causal SNP, SNP_c, and that of any other SNP and f is the MAF of the causal SNP. All SNPs are coded as 0, 1 and 2 according to the number of minor alleles. To simulate an LD pattern in which the correlation between SNP i and the causal SNP is dependent on physical proximity we use the following equation to determine

the correlation, ρ_i, between a SNP_i and SNP_c,

$$\rho_i = \rho_{max}\left(1 - 2\frac{|i-c|}{k}\right) \tag{6}$$

where position $i = 1 \ldots k$ and c is the position of the causal SNP among k SNPs.

Patterns I, III and IV have ρ_{max} of 0.9 and Pattern II has a lower maximum correlation of 0.5. Pattern III is designed to produce a flat haplotype block within the region, hence Equation 6 is conditioned such that if $\rho_i > 0.5$ then ρ_i is set to ρ_{max}. Under pattern IV each SNP obtains its correlation to the closest causal SNP according to Equation 6.

The genotypes are simulated by creating a latent variable $Z_c \sim N(0,1)$ for SNP_c and then for each of the other SNPs,

$$Z_i = \frac{\rho_i Z_c + (1-\rho_i)\varepsilon_i}{\sqrt{(\rho_i^2 + (1-\rho_i)^2)}} \tag{7}$$

where $\varepsilon_i \sim N(0,1)$ is drawn randomly.

The genotypes for SNP i are created by specifying a MAF, f and then coding the genotype as 0 if $\Phi(Z_i) < (1-f)^2$, 1 if $(1-f)^2 < \Phi(Z_i) < 1-f^2$ and 2 otherwise, where $\Phi()$ is the standard normal integral.

3.2 GENOME simulator

The genome simulator, GENOME [47], was used to provide a more realistic but less controllable set of genetic variants. GENOME applies the coalescent-based approach [43] to simulate genome data using the Wright-Fisher neutral model [24]. The algorithm uses realistic recombination rates, genealogy tree and haplotype blocks. GENOME was used to simulate 10,000 haplotypes of the size of an average protein-coding gene (53.6×10^3bp) with 200 SNPs [63] and 5 recombination points [30]. The genotypes of individuals were simulated by selecting 2 out of the 10,000 haplotypes and combining them, until the required genotypes for 2,000 individuals were obtained.

3.3 The exposure and outcome

The simulated exposure and outcome were selected to mimic the relationship between birth weight and type II diabetes as measured by the level of fasting glucose [71]. Each simulated dataset consisted of the genotype of a causal SNP (SNP_c), risk factor (X), disease outcome (Y), and unmeasured confounding (U) for 2,000 individuals. The causal SNP explained 2% of the variation in X [37] and X explained 6% of the variation in Y. X and Y were distributed $N(3.3, 0.59^2)$ [Office for National Statistics] and $N(5.47, 1.32^2)$ [23] respectively. U was drawn from a $N(0,1)$ distribution. The following equations describe the relationship between SNP_c, X, Y and

U;

$$X_i = \alpha_0 + \alpha_1 \text{SNP}_{ci} + \alpha_2 U_i + \varepsilon_{xi} \tag{8}$$

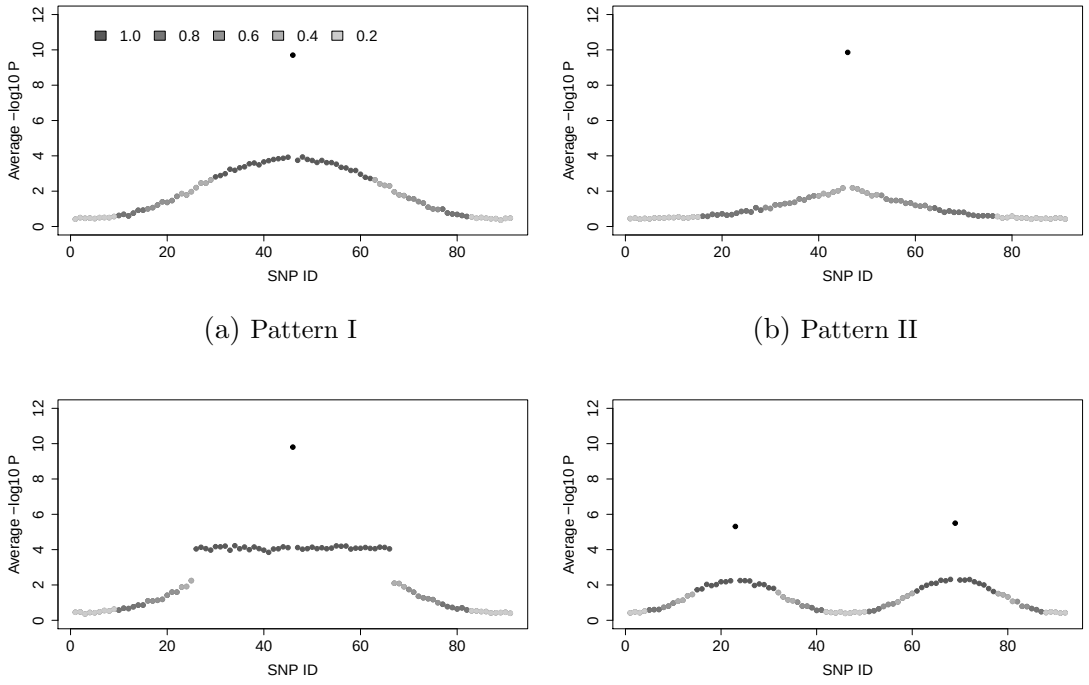
$$Y_i = \beta_0 + \beta_{XY} X_i + \beta_2 U_i + \varepsilon_{yi} \tag{9}$$

where ε_{xi} and ε_{yi} are independent random errors with distributions of $N(0,1)$, $i = 1, \ldots, n$ and n is the number of individuals. As the causal SNP explains 2% of the variance in X, the remaining 98% was divided equally between U and ε and their regression coefficients are calculated accordingly.

In LD Pattern IV there were two causal SNPs, SNP_{c1} and SNP_{c2}, so the regression equation for X became;

$$X_i = \alpha_0 + \alpha_1 \text{SNP}_{c1i} + \alpha_2 \text{SNP}_{c2i} + \alpha_3 U_i + \varepsilon_{xi} \tag{10}$$

where each causal SNP explained 1% of the variation in X.



(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 1: The regional association plots for the four patterns with 90 SNPs. On the x-axis is SNP ID representing chromosome position, and on the y-axis is -$\log_{10} P$, where P is the mean p-value, over 200 simulations, from the regression of X on each SNP individually. Colour coding (from light to dark grey) denotes strong to weak correlation with the causal SNP; see also the legend within the plot. The black dot is the causal SNP

Figure 1 shows regional plots of the average values of -log10(p-value) for Patterns I, II, III and IV, when the p-values are calculated separately for each SNP. The

strength of correlation decreases with the distance from the causal SNP and this is reflected in the decline in average statistical significance.

4 Convergence and mixing

When there are k potential instruments, there are 2^k possible models linking the selected instruments to the exposure. Adequately covering the huge space of potential models is a challenge and a poorly tuned algorithm will mix badly and so be slow to converge. An important aspect of any variable selection algorithm is the choice of priors, because when an instrument is excluded from the current model, it is considered for inclusion at a later point with a coefficient that will be dependent on those priors. The priors in *ivbma* are chosen to be relatively vague and they cannot be altered by the user. For this reason it is important to consider whether the choices imposed by *ivbma* work well for Mendelian randomisation with many instruments. O'Hara *et al.* [52] gives a general discussion of the use of priors to improve mixing and increase the speed of convergence.

Karl *et al.* [41] suggested that, for most problems, 50,000 iterations and a 10,000 burn-in would be sufficient to reach convergence with *ivbma*. However, they have only looked at scenarios seen in the econometric literature. To investigate convergence in the context of MR we considered a range of scenarios and analysed them with *ivbma* using 5 chains with random starting values, each chain having 50,000 iterations and a 10,000 burn-in. We then analysed the same data with a single chain of length 500,000 with a burn-in of 250,000.

In the first scenario we varied the number of potential instruments and considered each of 10, 30, 60 and 90. The causal SNP had MAF of 0.05 but was excluded from the analysis and the correlated non-causal SNPs had MAF varying randomly between 0.1 to 0.5. The second scenario was similar except there were just 10 instruments and the MAF of the non-causal SNPs was fixed at 0.1. A third scenario also considered 10 instruments but with negative correlation between X and Y, and with a stronger confounder.

We assessed mixing by looking at the trace plots and assessed convergence by the 5 chains with different initial values [48]. For each analysis of the same dataset we monitored the mean causal effect estimate, 95% credible interval, the probability that X was included in the second regression model (i.e. the probability of a causal effect) and the total visited probability for the 10 most visited models in the long chain [31].

The full results of this experiment are given in Supplement Table S.1. To summarise the results we calculated the mean absolute deviation (MAD) between the estimate of the causal effect, β_{XY}, from the 5 short chains and the estimate from the long chain. In the first scenario in which the number of instruments varied, the MADs were 0.013, 0.014, 0.025, 0.019. The agreement between chains deteriorated only slightly as the number of instruments increased and averaged at about 7% of the true value of the causal parameter. The low fixed MAF scenario had a MAD

of 0.015 and in scenario three the MADs were 0.027 and 0.013 for negative and stronger confounding effects, respectively.

Convergence and mixing of the shorter chains did not vary greatly across the scenarios that we considered. We decided that the short chains were sufficiently accurate for the MR scenarios to show strong patterns in the scenarios that we wanted to consider and we decided to use 50,000 iterations and a 10,000 burn-in for all analyses.

5 Selection of instruments

In the situation in which *ivbma* has many correlated genetic instruments to choose from we wanted to look for patterns in the way that SNPs were selected so we simulated three scenarios as shown in Table 1. Each dataset consisted of 10 non-causal SNPs genotyped on 2,000 individuals. The SNPs had LD of Pattern I with the causal SNP located in the centre of the range between SNP_5 and SNP_6, but the causal SNP was not included in the analysis. The pattern of MAF was varied; in the common scenario the MAF was 0.45, in the low scenario it was 0.1 and in the variable scenario it varied randomly between 0.1 and 0.5. Each scenario was simulated 200 times and the results were averaged across the simulations.

When *ivbma* analyses the data, it averages over different combinations of the instruments and over models that conclude that there is a causal link between X and Y with a non-zero coefficient β_{XY} or that there is no causal link, in which case β_{XY} is zero. Each dataset provides its own probability of a causal relationship by the proportion of the chain for which β_{XY} is non-zero. Table 1 shows that in the common MAF scenario the probability on a causal relationship between X and Y was about 0.69. So there is a 31% chance that X is not causally related to Y, which explains why the average estimate of β_{XY} is downwardly biased. The sum of the probabilities of inclusion for the ten SNPs in the common MAF scenario is 1.67 indicating that on average the MR used 1.67 SNPs. SNP selection strongly favoured those SNPs that were most highly correlated with the causal SNP. In the low and variable MAF scenarios, the patterns were similar although for the case of variable MAF the tendency to select the highly correlated SNPs is less strong, since sometimes these will have low MAF.

Figure 2 shows the posterior distributions of β_{XY} for two simulated datasets; the one on the left is typical of the situation in which there is a high probability of a causal relationship between X and Y and the one on the right illustrates the situation in which the evidence of causality is less strong.

Table 1: Average performance over 200 datasets between three different MAFs; $\hat{\beta}_{XY}$ is the causal effect estimate (true value, 0.2449) and SE is its standard error. p(X) is the probability that X is included in the second regression. Correlation is with the causal SNP. $\hat{\beta}_{ZX}$ is the estimated association of the SNP with X. p(SNP) is the probability of being included as an instrument.

MAF	Mean $\hat{\beta}_{XY}$ (SE)	p(X)	SNP	Correlation	Mean $\hat{\beta}_{ZX}$	p(SNP)
Com	0.190 (0.018)	0.693	1	0.089	0.000	0.044
			2	0.308	0.001	0.045
			3	0.568	0.003	0.074
			4	0.801	0.012	0.148
			5	0.941	0.076	0.541
			6	0.941	0.067	0.490
			7	0.801	0.011	0.146
			8	0.570	0.003	0.082
			9	0.310	0.002	0.059
			10	0.088	0.000	0.038
Low	0.167 (0.022)	0.561	1	0.089	0.000	0.072
			2	0.307	0.000	0.080
			3	0.570	0.006	0.102
			4	0.801	0.027	0.216
			5	0.942	0.078	0.446
			6	0.942	0.103	0.556
			7	0.801	0.021	0.195
			8	0.570	0.008	0.121
			9	0.308	0.000	0.070
			10	0.089	-0.001	0.089
Var	0.171 (0.022)	0.557	1	0.089	-0.001	0.063
			2	0.310	0.000	0.064
			3	0.568	0.003	0.087
			4	0.801	0.016	0.187
			5	0.941	0.047	0.381
			6	0.941	0.046	0.378
			7	0.801	0.017	0.183
			8	0.569	0.002	0.077

Continued on next page

Table 1 – *Continued from previous page*

MAF	Mean $\hat{\beta}_{XY}$ (SE)	p(X)	SNP	Correlation	Mean $\hat{\beta}_{ZX}$	p(SNP)
			9	0.308	-0.001	0.073
			10	0.090	-0.003	0.072

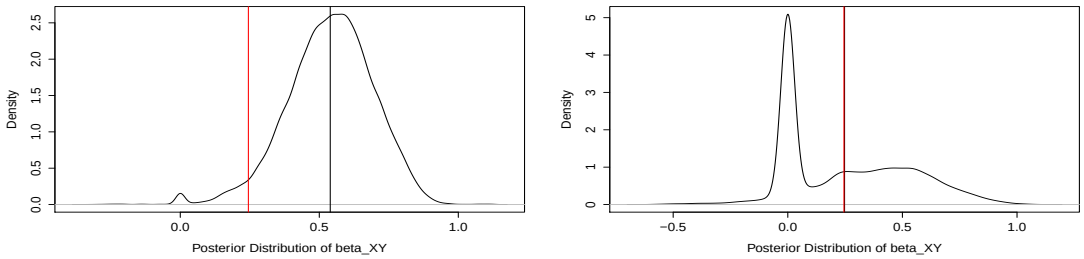


Figure 2: Posterior distribution of causal effect (β_{XY}) for example datasets. In the left hand example, the probability that X is causal is high (0.99) while in the right hand example causality is less certain (0.64). The red and black vertical lines represent the true causal effect (0.24) and posterior mean, respectively.

6 Comparison with other estimators

In this section we compare *ivbma* with two classical approaches, 2SLS and LIML, and with a recently suggested approach based on penalised regression. The first two simulation experiments use the controlled patterns of LD introduced in subsection 3.1; experiment 1 investigates the impact of MAF on relative performance and experiment 2 investigates the impact of LD pattern. The third experiment uses GENOME to simulate realistic patterns of LD and MAF as described in subsection 3.2. Each scenario was repeated 200 times and each dataset was analysed using *ivbma* with 50,000 MCMC iterations and burn-in of 10,000.

As recommended by Burton *et al.* [13], bias, root mean squared error (RMSE) and coverage were monitored. LIML is known to produce occasional outliers in the presence of many weak instruments, that is extreme estimates of the causal effect size, [15]. To compare performance without the effect of outliers, bias and RMSE measures were winsorised when combining over the 200 simulated datasets and the percentage of outliers was noted. Winsorisation reduces the effect of outliers by replacing the highest and lowest 20% of the estimates by the 20% and 80% quantiles [72]. Outliers were defined as in a box plot, that is to say, more than 1.5 times the interquartile range from the upper or lower quartile.

The R package *sisVIVE* was used to implement the penalised regression based method of Kang *et al.*[40]. As recommended by the authors, we used 10-fold cross-

validation for the selection of the penalty parameter. *sisVIVE* does not provide the standard errors for the point estimate and so coverage is not reported.

6.1 Effect of MAF

In experiment 1, the correlation between SNPs had Pattern I; for the variable case, the causal SNP had a MAF of 0.5 and non-causal MAFs were randomly generated between 0.1 and 0.5. In the common case, the causal and non-causal SNPs had MAFs of 0.45 and 0.5 respectively. In the low MAF scenario, the MAFs were 0.05 and 0.1 respectively. The results are shown in Table 2.

Although commonly used when there are few instruments, 2SLS, is not a practical method for many dependent instruments, even in the common MAF scenario the bias increases with the number of available instruments, coverage deteriorates sharply and the RMSE is worse than for any of the other methods. Performance with low or variable MAF shows the same patterns but is worse. *sisVIVE* also shows deteriorating performance as the number of available instruments increases and is similar to 2SLS. LIML and *ivbma* perform much better, performance is generally more stable as the number of potential instruments increases, bias is less, RMSE is smaller and coverage is better. Of the two, *ivbma* performs slightly better than LIML with better RMSE and more stable coverage, especially with low or variable MAF.

ivbma has the advantage of a prior that restricts the range of possible causal effect estimates and as a result outliers are rare regardless of the MAF or number of instruments. Unexpectedly, for 10 and 30 instruments with variable MAF, *ivbma* produced more outliers than 2SLS; these came about because of uncertainty in the existence of any causal association between X and Y, see Supplement Figure S.5b.

Table 2: Average performance with different patterns of minor allele frequency (MAF) for 2SLS, LIML, *ivbma* and *sisVIVE*. Inst. is the number of instruments. Both bias and RMSE are Winsorised. Outlier is the percentage of extreme estimates.

MAF	Inst.	2SLS		LIML		*sisVIVE*		*ivbma*	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Com	10	0.089	0.129	-0.041	0.147	0.144	0.167	-0.058	0.115
	30	0.217	0.228	-0.015	0.137	0.327	0.253	-0.059	0.113
	60	0.309	0.315	0.002	0.162	0.358	0.331	-0.060	0.110
	90	0.338	0.343	0.187	-0.023	0.300	0.361	-0.024	0.109
Low	10	0.146	0.176	0.005	0.130	0.165	0.192	-0.098	0.141
	30	0.248	0.261	-0.049	0.184	0.258	0.267	-0.068	0.132
	60	0.340	0.344	-0.007	0.208	0.339	0.345	-0.006	0.119

Continued on next page

Table 2 – *Continued from previous page*

MAF	Inst.	2SLS		LIML		*sisVIVE*		*ivbma*	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	90	0.369	0.372	-0.072	0.302	0.365	0.369	0.021	0.119
Var	10	0.164	0.204	-0.011	0.182	0.186	0.217	-0.098	0.136
	30	0.285	0.300	-0.022	0.228	0.300	0.310	-0.075	0.137
	60	0.342	0.350	-0.043	0.261	0.366	0.370	-0.044	0.133
	90	0.390	0.393	0.054	0.279	0.384	0.387	0.042	0.151
		Outlier	Coverage	Outlier	Coverage	Outlier	Coverage	Outlier	Coverage
Com	10	5.0	86.5	3.0	92.0	2.5	-	0.0	87.5
	30	1.0	47.5	1.5	88.0	0.0	-	0.0	88.0
	60	0.5	10.0	4.5	83.0	2.5	-	0.0	90.0
	90	0.0	2.0	7.0	79.5	2.0	-	0.0	90.5
Low	10	1.0	86.0	5.5	96.0	1.0	-	3.5	96.5
	30	1.0	45.0	5.0	91.5	0.5	-	0.0	93.0
	60	1.0	7.0	4.0	84.5	3.0	-	0.0	94.5
	90	2.0	0.5	7.5	76.0	0.5	-	0.0	92.0
Var	10	0.5	82.0	5.0	94.5	1.5	-	1.5	96.5
	30	0.5	41.5	8.0	89.5	1.5	-	1.0	95.0
	60	0.5	10.0	7.0	79.5	0.5	-	0.5	92.5
	90	1.5	0.5	10.0	72.5	1.5	-	0.0	91.0

6.2 Four patterns of LD

In experiment 2, the MAF of causal SNP was 0.5 and non-causal MAFs varied between 0.1 and 0.5. LD patterns I, II, III and IV were all considered. The results are presented in Table 3.

When the number of instruments is small, *ivbma* is more biased than LIML for all of the LD patterns, reflecting the fact that *ivbma* is uncertain about the causality of X. Under patterns III and IV with 90 instruments, the bias with *ivbma* becomes positive rather than negative as the causality of X becomes clearer (Supplement Figure S.6). LIML sometimes has too many outliers for 20% Winsorisation to remove. For Pattern II there are hardly any improvements in performance as the number of potential instruments increases reflecting the fact that none of them was strongly correlated with the causal SNP. However, the Winsorised RMSE from LIML are

large in comparison to *ivbma*.

Table 3: Average performance with LD Patterns I, II, III and IV for 2SLS, LIML, *ivbma* and *sisVIVE*. Inst. is the number of instruments. Both bias and RMSE are Winsorised. Outlier is the percentage of extreme estimates.

LD	Inst.	2SLS		LIML		*sisVIVE*		*ivbma*	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
I	10	0.164	0.204	-0.011	0.182	0.186	0.217	-0.098	0.136
	30	0.285	0.300	-0.022	0.228	0.300	0.310	-0.075	0.137
	60	0.342	0.350	-0.043	0.261	0.366	0.370	-0.044	0.133
	90	0.390	0.393	0.054	0.279	0.384	0.387	0.042	0.151
II	10	0.232	0.274	-0.022	0.307	0.271	0.298	-0.059	0.131
	30	0.335	0.345	0.006	0.283	0.337	0.351	-0.051	0.142
	60	0.374	0.381	-0.001	0.319	0.385	0.390	-0.030	0.120
	90	0.399	0.402	-0.039	0.372	0.405	0.408	0.003	0.145
III	10	0.158	0.191	-0.038	0.200	0.205	0.231	-0.102	0.142
	30	0.266	0.281	-0.010	0.194	0.274	0.286	-0.061	0.137
	60	0.339	0.345	0.008	0.209	0.348	0.352	-0.027	0.131
	90	0.369	0.372	-0.058	0.252	0.392	0.396	0.026	0.142
IV	10	0.189	0.227	-0.053	0.245	0.479	0.479	-0.106	0.144
	30	0.306	0.315	0.002	0.241	0.481	0.481	-0.064	0.136
	60	0.372	0.378	0.025	0.257	0.480	0.480	-0.007	0.131
	90	0.391	0.393	-0.035	0.309	0.480	0.480	0.026	0.131
		Outlier	Coverage	Outlier	Coverage	Outlier	Coverage	Outlier	Coverage
I	10	0.5	82.0	5.0	94.5	1.5	-	1.5	96.5
	30	0.5	41.5	8.0	89.5	1.5	-	1.0	95.0
	60	0.5	10.0	7.0	79.5	0.5	-	0.5	92.5
	90	1.5	0.5	10.0	72.5	1.5	-	0.0	91.0
II	10	1.5	77.0	11.0	92.0	2.5	-	1.0	97.0
	30	1.0	30.0	9.5	83.5	0.0	-	0.0	97.5
	60	1.5	4.5	14.0	80.0	0.0	-	0.5	96.5
	90	0.5	0.0	6.5	70.0	1.0	-	0.0	91.5

Continued on next page

Table 3 – *Continued from previous page*

LD	Inst.	2SLS		LIML		*sisVIVE*		*ivbma*	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
III	10	0.5	82.5	5.0	93.5	0.5	-	3.0	92.5
	30	0.0	43.0	4.5	87.5	0.5	-	1.0	94.5
	60	0.0	5.0	6.5	82.5	1.5	-	0.0	93.0
	90	1.5	1.0	10.5	75.5	1.0	-	0.0	89.0
IV	10	2.0	79.5	4.5	91.5	1.0	-	4.0	95.5
	30	2.5	33.0	8.0	84.5	0.0	-	0.5	93.5
	60	0.5	5.5	10.0	79.0	0.0	-	0.0	94.5
	90	2.5	0.5	9.5	76.5	1.0	-	0.0	89.5

6.3 GENOME

GENOME was used to simulate realistic patterns of LD for a region containing approximately 200 SNPs. The causal SNP was randomly selected from those SNPs with a MAF below 0.1 and all common SNPs with a MAF of at least 0.1 were retained for the analysis. The simulation was repeated 200 times and the average number of common SNPs available for analysis was 41 (range 3 to 87). In one simulation there were no common SNPs and this dataset was discarded. The 199 datasets were analysed by 2SLS, LIML, *sisVIVE* and *ivbma* and the results are given in Table 4. *ivbma* performed best on all scales with the lowest Winsorised bias, the smallest RMSE, the fewest outlying estimates and coverage that was closest to 95%.

Table 4: Average performance for 2SLS, LIML, *ivbma* and *sisVIVE* with GENOME simulated genetic instruments. $\hat{\beta}_{XY}$ is the causal effect estimate (true value 0.2449). Both bias and RMSE are Winsorised. Outlier is the percentage of extreme estimates.

	Mean $\hat{\beta}_{XY}$	Bias	RMSE	Outlier	Coverage
2SLS	0.417	0.164	0.192	1.0	73.9
LIML	0.755	-0.282	0.902	8.0	10.6
sisVIVE	0.386	0.132	0.156	4.0	-
ivbma	0.196	-0.069	0.132	0.0	94.5

7 Robustness

In this section we investigate the robustness of *ivbma* to non-normality in the confounding between X and Y, and to a violation of assumption (3) for Mendelian

randomisation.

All experiments were based on simulated data for a MR on 2,000 individuals. The LD pattern I of Figure 1 was used, so there was a single causal SNP. The causal SNP was not included in the analysis. The non-causal SNPs had MAFs randomly chosen to lie between 0.1 and 0.5, and the causal SNP had a MAF of 0.45. Performance was assessed with 10, 30, 60 and 90 instruments averaged over 200 random datasets each analysed using *ivbma*.

Table 5: Average performance with different number of SNPs and variable MAF, when the errors have various distributions and in the presence of invalid instruments. Inst. is the number of Instruments. $\hat{\beta}_{XY}$ is the causal effect estimate (true value 0.2449) and SE is the standard error.

Distribution	Invalid	Inst.	Mean $\hat{\beta}_{XY}$	SE	Bias	RMSE	Outlier	Coverage
Normal	**0**	10	0.171	0.011	-0.074	0.176	1.5	96.5
		30	0.181	0.012	-0.064	0.175	1.0	95.0
		60	0.215	0.012	-0.030	0.174	0.5	92.5
		90	0.291	0.013	0.046	0.190	0.0	91.0
Uniform	**0**	10	0.181	0.009	-0.064	0.140	0.0	85.5
		30	0.192	0.008	-0.053	0.128	0.0	91.0
		60	0.208	0.008	-0.037	0.124	0.0	89.5
		90	0.237	0.009	-0.008	0.121	0.0	91.0
Student's t	**0**	10	0.194	0.009	-0.052	0.140	0.0	87.5
		30	0.207	0.009	-0.038	0.135	0.0	88.0
		60	0.210	0.009	-0.035	0.131	0.0	88.0
		90	0.220	0.008	-0.025	0.122	0.0	92.0
Normal	**1**	10	0.345	0.009	0.100	0.162	0.1	73.0
		30	0.335	0.009	0.090	0.154	0.0	74.0
		60	0.336	0.009	0.091	0.155	0.0	78.5
		90	0.340	0.008	0.095	0.151	0.1	74.0
Normal	**10%**	10	0.332	0.009	0.087	0.158	0.0	77.0
		30	0.415	0.008	0.170	0.204	0.0	58.5
		60	0.489	0.007	0.244	0.261	0.0	29.0
		90	0.554	0.007	0.309	0.324	0.0	10.0

7.1 Non-normal errors

As in previous sections the confounding between X and Y was first simulated using a standard normal distribution and the first block of Table 5 shows the average results.

To assess distributional robustness we then simulated confounding from a short-tailed and a long-tailed distribution with the same mean and variance, namely a Uniform distribution ($unif(-\sqrt{3}, \sqrt{3})$) and a Students t distribution with 4 degrees of freedom ($t(4)/\sqrt{2}$). The results are shown in the second and third blocks of Table 5. There is a slight reduction in the coverage when the data are simulated with non-normal confounding but no evidence of an effect on the mean estimate. The distributional assumptions do not seem to have an important effect on the results.

7.2 Invalid Instruments

Invalid instruments were defined as instruments that are directly associated with Y along a pathway that does not pass through X; this is sometimes known as horizontal pleiotropy. Two scenarios were considered; only 1 invalid instrument and 10% of instruments invalid. The non-causal SNP(s) with the lowest correlation with the causal SNP were chosen as the invalid instrument(s). The invalid instrument(s) in total explained 0.1% of the variation in Y.

The results are shown in Table 5. In the presence of 1 invalid instrument, *ivbma* remains consistently biased regardless of the number of instruments and that bias is comparable in size to that seen with no invalid instrument but in the opposite direction. Coverage is much poorer. When 10% of the instruments are invalid the effects are similar in that both bias and coverage deteriorate but the effects increase with the number of instruments. It appears that *ivbma* is sensitive to the inclusion of invalid instruments particularly if the number of invalid instruments increases with the number of potential instruments.

8 GRAPHIC Study: *FTO* gene, body mass index and blood pressure

There are many epidemiological studies demonstrating the association between obesity and blood pressure, although the exact mechanism behind this relationship is unknown. As a consequence it is difficult to identify all of the confounders and causality is uncertain [21, 62, 74]. To illustrate the used of *ivbma* with real data we considered the causal effect of body mass index (BMI) on mean 24 hour systolic blood pressure (SBP) using instruments taken from the *FTO* gene.

8.1 Data

The data come from the Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) study [69]. This population-based cohort recruited 2037 white European participants from 520 nuclear families living in Leicestershire, UK. To avoid the complication of a family effect, only parents' data were analysed. Individuals were included if they had complete data for body mass index (BMI) and mean 24-hour systolic blood pressure (SBP).

8.2 Instruments

The *FTO* gene has been identified as being associated with BMI in GWAS [27]. 207 measured SNPs in the region of the *FTO* gene were selected. A SNP was included in the analyses if (1) it had less than 1% of missing data, (2) the minor allele frequency was greater than 0.1 (3) it was in Hardy-Weinberg equilibrium (4) it was not in full LD with another SNP. After quality control there were 173 BMI-related SNPs that acted as the set of potential instruments. The genotypes of the SNPs were coded 0, 1, or 2 representing the number of BMI-increasing alleles.

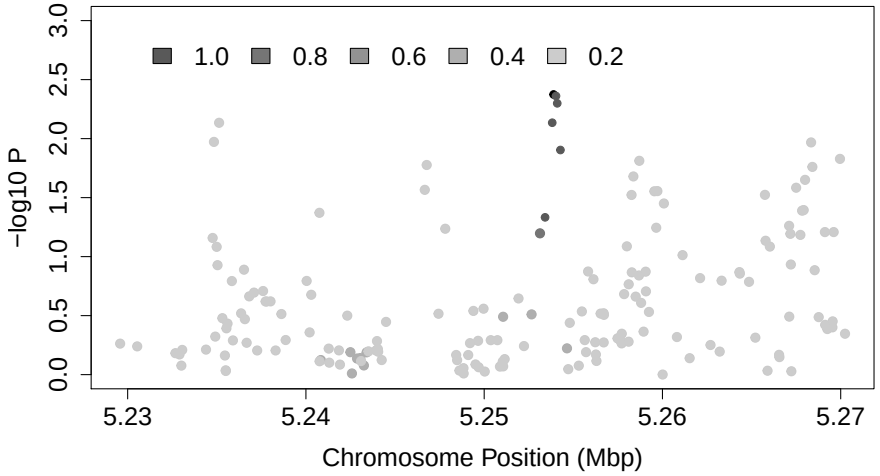


Figure 3: Regional association plots for BMI-related *FTO* variants. Regional P value plots where the p-value is from the regression of each SNP on BMI. On the x axis is SNP ID in the ascending order of chromosome position, and on the y axis is -$\log_{10} P$. Colour coding (from light to dark grey) denotes LD information; see also the legend within the plot.

8.3 Results

After quality control, there were 1026 unrelated-individuals with complete records of BMI and mean 24-hr SBP. Table 6 summarises the characteristics from the subjects.

Table 6: GRAPHIC study unrelated-individuals characteristics, N=1028

	N	Mean	Standard deviation
Gender (male)	1028(514)	-	-
Age (years)	1028	52.7	4.6
BMI (kg/m^2)	1028	27.4	4.3
Mean 24-hr SBP (mm Hg)	1026	120.6	12.0

All the SNPs from GRAPHIC would be considered as weak instruments, as their individual F-statistics from the association with BMI were less than 10 [61]. Even for the lead SNP, the F-statistic was 8 and explained approximately 1% of the variation in BMI. Figure 3 shows that out of 173 SNPs, only 6 were strongly correlated with the lead SNP. There was a second SNP, independent of the lead SNP, that was also significant and which had a similar effect size for BMI as the lead SNP. This resembles Pattern IV in the simulation study but with weaker instruments.

As the R package *ivbma* does not have an option to change the prior distribution, we have standardised X and Y and Z, to ensure that the prior does not heavily influence the posterior distribution. For comparability to the other estimates, the coefficient and credible interval from *ivbma* shown in Table 7 have been transformed back to the original scale.

For the non-Bayesian methods, Table 7 gives the estimated coefficient of the causal effect, its 95% confidence interval and p-value. For the Bayesian approach adopted by *ivbma* we give the posterior mean estimate of the coefficient, its 95% credible interval and the probability of a causal association as measured by the probability that BMI is included in the regression equation for mean SBP. The analyses give slightly different pictures of the BMI-SBP association. 2SLS, *sisVIVE* and *ivbma* all agree that there is a positive association and there is overlap of their confidence regions. The estimates from 2SLS and *sisVIVE* are similar to that which would have been obtained by OLS regression, 0.90 (95% CI: 0.74, 1.07). LIML estimated the effect of BMI in the opposite direction, but the confidence interval is very wide.

Table 7: The effect of BMI (kg/m^2) on SBP (mm Hg), where N=1026 and there are 173 potential instruments. P(causal), posterior probability of BMI and SBP having a causal relationship.

Method	Coefficient (95% Confidence Interval)	p-value
2SLS	0.86 (0.47, 1.25)	<0.001
LIML	-5.72 (-14.94, 3.50)	0.2246
sisVIVE	0.92 (-,-)	-
	Coefficient (95% Credible Interval)	P(causal)
ivbma	1.21 (-0.08, 2.58)	0.98

8.4 Conclusion

2SLS is known to be biased towards the OLS estimate with weak instruments [61] and *sisVIVE* also gave a similar causal effect estimate to OLS. As seen from Section 6, LIML usually gives better estimates with many instruments but will sometimes give extreme estimates when the dataset only has weak instruments. This is the case for GRAPHIC study; the highest F-statistic for all the SNPs was 8. The

estimated coefficient from *ivbma* is larger than that provided by the other methods at 1.21 mmHg per unit increase in BMI. If correct, this would suggest that BMI is a more serious public health issue, at least over the range of BMI represented in this study. However, the associated credible interval indicates a lack of precision around this estimate. In 2009, Timpson *et al.* [68] performed a MR analysis of the Copenhagen General Population study of 36,851 participants and showed that SBP increased by 3.85 (95% CI:1.88, 5.83) per 10% increase in BMI, using *FTO* and *MC4R* as the two instruments. Using a meta-analysis of 30 studies on the effect of the *FTO* genotype on SBP and BMI, the MR analysis resulted in an increase of 0.89 (95% CI: 0.48, 1.31) SBP for a 1-unit of increase in BMI [25]. Holmes *et al.* [36] performed a MR using individual-level data from 6 studies (N=30,136) and demonstrated that SBP increased by 0.70 (95% CI: 0.24, 1.16) per unit increase in BMI. In that study the instrument was a genetic score from 14 SNPs, weighted by the coefficients from a discovery study [34].

9 Discussion

Instrumental variable analysis with many dependent, weak instruments represents a difficult problem for any method of analysis, but through simulations we have demonstrated that *ivbma* performs better many of the alternatives. Our analyses have relied heavily on the R package *ivbma*. Similar estimates can be obtained from OpenBUGS [59] but the convergence of that algorithm is much slower and it is not practical for use in a simulation study. However, OpenBUGS does have the advantage of allowing the use of different prior distributions.

In our simulations with controlled patterns of LD, *ivbma* consistently out-performed the other methods that we tried regardless of the pattern of LD or the MAF or the number of instruments. The simulations with realistic patterns of LD also strongly favoured *ivbma* and the study with real data provided a posterior mean estimate that is similar to estimates from other epidemiological studies. Further evidence of the method's performance with real data is required.

When the causal relationship between exposure, X, and outcome, Y, is uncertain, IVBMA will produce a posterior distribution for the causal effect estimate that has a bimodal shape with one peak at 0 for models without X and another peak at the mean estimates from models which include X (as shown in Figure 2). As a result the posterior mean will be pulled towards zero and to get a true impression of the results it is important to plot the posterior distribution of the causal effect. To help judge the results of BMA, Kass and Raftery [42] suggested a posterior inclusion probability of <0.5 suggests no effect of the explanatory variable on the outcome, 0.5-0.75 gives weak evidence for an effect, 0.75-0.95 gives positive evidence and >0.95 gives strong evidence. This scale might be adopted when deciding on causality.

The algorithm incorporated into *ivbma* [41] has been shown to be more computationally efficient than Two-stage BMA (2BMA) [46], a method similar to 2SLS, and it has been shown to have faster convergence and better mixing that the al-

gorithm of Koop *et al.*'s [44] approach. However, Karl and Lenkoski [41] did find that convergence of the 95% credible interval from *ivbma* deteriorates as number of potential covariates in the regression increases.

Our simulations have shown that in the presence of invalid instruments, *ivbma* performs poorly. Bias and the proportion of outliers increased and coverage decreased with increasing numbers of instruments. This may be because these invalid instruments also explain X, from X's association with Y. Hence, IVBMA cannot distinguish between the direct effect on Y or the indirect effect through X on Y. Furthermore, the more invalid instruments that are included with similar causal effect estimates, the more certain *ivbma* is of their effect being genuine and therefore these are given more weight in the averaged causal effect estimation. This is a serious practical concern since it will be almost impossible to be certain that all potential instruments are valid. A possible solution might be to implement a version of BMA with the genetic variants included in the regression equation for Y but this needs further investigation. Invalid instruments are also a major concern in the classical MR setting where several methods to account for this have been developed, each with their own estimation assumptions, and suitable for both summary data [5, 6] and individual-level data [40, 73].

One of the limitations of this study is that we have not compared *ivbma* with all of the approaches for many instruments [2, 3, 14, 16]. However, the many instruments literature concentrates on strong instruments and may therefore not be relevant to most Mendelian randomisation settings. IVBMA is particularly appealing in the context of many weak instruments because, by definition, model selection will be difficult and model averaging allows us to incorporate that uncertainty.

As well as producing point estimates of the causal effect with good repeated sampling properties, the Bayesian nature of IVBMA offers two other advantages. Firstly, it provides a direct estimate of the probability that the exposure is causal in terms of the probability that the exposure is included in the regression equation for the outcome. Secondly, it offers the potential for including biological knowledge in the form of informative prior distribution. However it is not easy to quantify prior knowledge and the IVBMA analysis would have to be programmed in software such as OpenBUGS or Stan if the priors were to be changed. The use of informative priors could represent how to make use of many weak instruments in small studies.

In the general Bayesian statistical literature the focus is on priors that improve the mixing and speed of convergence; O'Hara *et al.* [52] provide an insightful review of priors and samplers for Bayesian variable selection but when the information is weak there may also be benefits in adopting informative prior distributions. The assumptions of Mendelian randomisation ought to be validated by biological knowledge [8, 54] and in a Bayesian approach this same knowledge could be incorporated into the priors.

In the related field of determining genetic association, Fridley *et al.* [28], considered the use of some of the Bayesian variable selection algorithms described in O'Hara *et al.* [52] with SNPs as potential predictors. However, there was very little discussion on how to quantify priors on the model space. Other studies

have considered SNP prioritisation to increase statistical power in GWAS analysis [29, 49, 67] effectively applying weights in the gene-exposure regression. Such weights might be based on previous GWAS significance, perhaps from GWAS Central at *www.gwascentral.org*. Alternatively weights might vary by haplotype block or SNPs within protein coding genes might be given more weight [49]. Incorporating prior information for correlated variables can be complicated [38] especially in a variable selection problem where the prior might dependent on which other variables are included in the current model.

Decisions on the prior for the second regression in an instrumental variable analysis is problem specific, our analyses and those of Koop *et al.* [44] and Karl and Lenkoski [41] all allow X to come in and out of the regression model for Y and hence the causal effect can be zero. If the investigator is sure of the causal relationship between X and Y then that prior knowledge could be used to place a prior on the causal effect estimate that excludes zero and which perhaps dictates the direction of the effect.

Jones *et al.* [39] have found that an instrumental variable analysis is robust to the prior on the covariance matrix, as the model does not directly estimate the causal effect from the covariance matrix. Nevertheless, the prior on the covariance matrix does effect the precision of the causal effect estimate. BMA accounts for model uncertainty in the causal effect estimate, but without an informative prior on the covariance matrix, the precision of the causal effect could be very wide. The amount of confounding between X and Y might be assessed from previous clinical trials and epidemiological studies: when the coefficient decreases with adjustment for confounding this implies presence of positive confounding and if it increases this suggests negative confounding [9].

An important practical note is that the R package *ivbma* does not allow the user to change the priors, i.e. all the coefficients and intercepts are given a $N(0, 1)$ prior distributions. Because of this, we standardised the data when analysing the BMI-SBP study. This dataset is comparatively small and without transformation the prior would dominate the analysis and force the intercept in the second regression to be close to zero. As a consequence the causal estimate would have been artificially raised. Alternatively, the need for transformation can be avoided by running the IVBMA algorithm with vaguer priors in OpenBUGS [59], although the required computation time would be much greater.

In the future is would be really beneficial if the *ivbma* package were modified to allow greater flexibility in the choice of the prior. As well as having normal priors with much larger variances, it would be helpful to incorporate a g-prior in the first stage regression of MR, as this has been shown to move between models more efficiently for highly correlated predictors [4]. With the growth of GWAS, there has been a large effort to develop MR methods that can be used with summary data [10, 11, 55]. One approach to using summary data in IVBMA might be based on the algorithm from joint analysis of marginal summary statistics (JAM) [50]; this is a Bayesian variable selection that uses GWAS summary data to select the most likely causal SNP.

Bayesian analysis is never a black-box solution to data analysis but in the context of Mendelian randomisation where there are potentially many weak dependent instruments, IVBMA offers an interesting alternative method of analysis. Our results suggest that it is good alternative to more more established methods, although, like all Mendelian randomisation methods, the validity of the analysis is heavily dependent on strong assumptions about the instruments.

References

[1] M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261, 2004.

[2] P. A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

[3] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[4] L. Bottolo, S. Richardson, et al. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

[5] J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

[6] J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

[7] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

[8] S. Burgess and S. Thompson. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine*, 31(15):1582–1600, 2012. doi: 10.1002/sim.4498.

[9] S. Burgess, S. G. Thompson, C. R. P. C. H. D. G. Collaboration, et al. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764, Jun 2011.

[10] S. Burgess, A. Butterworth, and S. G. Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665, Nov 2013. doi: 10.1002/gepi.21758.

[11] S. Burgess, F. Dudbridge, and S. G. Thompson. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, 2015.

[12] S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, page 0962280215597579, 2015.

[13] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, Dec 2006. doi: 10.1002/sim.2673.

[14] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306, 2004.

[15] J. C. Chao and N. R. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

[16] V. Chernozhukov, C. Hansen, and M. Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *The American Economic Review*, 105(5):486–490, 2015.

[17] Z. Dastani, T. Johnson, F. Kronenberg, C. P. Nelson, T. L. Assimes, W. März, J. B. Richards, C. Consortium, A. Consortium, et al. The shared allelic architecture of adiponectin levels and coronary artery disease. *Atherosclerosis*, 229 (1):145–148, 2013.

[18] G. Davey Smith and S. Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, Feb 2003.

[19] G. Davey Smith and S. Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, Feb 2004. doi: 10.1093/ije/dyh132.

[20] V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16 (4):309–330, 2007.

[21] W. B. Drøyvold, K. Midthjell, T. I. L. Nilsen, and J. Holmen. Change in body mass index and its impact on blood pressure: a prospective population study. *International Journal of Obesity*, 29(6):650–655, 2005.

[22] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*, 9(3):e1003348, 2013.

[23] M. Elovainio, J. E. Ferrie, A. Singh-Manoux, M. Shipley, G. D. Batty, J. Head, et al. Socioeconomic differences in cardiometabolic factors: social causation or health-related selection? Evidence from the Whitehall II Cohort Study, 1991-2004. *American Journal of Epidemiology*, 174(7):779–789, Oct 2011. doi: 10.1093/aje/kwr149.

[24] W. J. Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media, 2012.

[25] T. Fall, S. Hgg, R. Mgi, A. Ploner, K. Fischer, M. Horikoshi, and *et al* for the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLOS Medicine*, 10(6):e1001474, 2013.

[26] E. Fisher, N. Stefan, K. Saar, D. Drogan, M. B. Schulze, A. Fritsche, et al. Association of AHSG gene polymorphisms with fetuin-A plasma levels and cardiovascular diseases in the EPIC-Potsdam study. *Circulation: Genomic and Precision Medicine*, 2(6):607–613, Dec 2009. doi: 10.1161/CIRCGENETICS.109.870410.

[27] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826): 889–894, 2007.

[28] B. L. Fridley. Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology*, 33(1):27–37, 2009.

[29] S. Friedrichs, D. Malzahn, E. W. Pugh, M. Almeida, X. Q. Liu, and J. N. Bailey. Filtering genetic variants and placing informative priors based on putative biological function. *BMC Genetics*, 17(2):33, 2016.

[30] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, et al. The structure of haplotype blocks in the human genome. *Science*, 296 (5576):2225–2229, 2002.

[31] E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

[32] M. Gögele, C. Minelli, A. Thakkinstian, A. Yurkiewich, C. Pattaro, P. P. Pramstaller, J. Little, J. Attia, and J. R. Thompson. Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. *American Journal of Epidemiology*, 175(8):739–749, 2012.

[33] S. Greenland. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729, 2000.

[34] Y. Guo, M. B. Lanktree, K. C. Taylor, H. Hakonarson, L. A. Lange, B. J. Keating, and The IBC 50K SNP array BMI Consortium. Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals. *Human Molecular Genetics*, 22(1):184–201, 2012.

[35] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401, 1999.

[36] M. V. Holmes, L. A. Lange, T. Palmer, M. B. Lanktree, K. E. North, B. Almoguera, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *American Journal of Human Genetics*, 94(2):198–208, Feb 2014. doi: 10.1016/j.ajhg.2013.12.014.

[37] M. Horikoshi, H. Yaghootkar, D. O. Mook-Kanamori, U. Sovio, H. R. Taal, B. J. Hennig, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature Genetics*, 45(1):76–82, Jan 2013.

[38] F. Hoti and M. Sillanpää. Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits. *Heredity*, 97(1):4–18, 2006.

[39] E. M. Jones, J. R. Thompson, V. Didelez, and N. A. Sheehan. On the choice of parameterisation and priors for the Bayesian analyses of Mendelian randomisation studies. *Statistics in Medicine*, 31(14):1483–1501, Jun 2012. doi: 10.1002/sim.4499.

[40] H. Kang, A. Zhang, T. T. Cai, and D. S. Small. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

[41] A. Karl and A. Lenkoski. Instrumental variable Bayesian model averaging via conditional Bayes factors. *arXiv preprint arXiv:1202.5846*, 2012.

[42] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[43] J. F. C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[44] G. Koop, R. Leon-Gonzalez, and R. Strachan. Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics*, 171(2): 237–250, 2012. doi: 10.1016/j.jeconom.2012.06.005.

[45] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, Apr 2008. doi: 10.1002/sim.3034.

[46] A. Lenkoski, A. Karl, and A. Neudecker. *ivbma: Bayesian Instrumental Variable Estimation and Model Determination via Conditional Bayes Factors*, 2014. R package version 1.05.

[47] L. Liang, S. Zöllner, and G. R. Abecasis. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567, Jun 2007. doi: 10.1093/bioinformatics/btm138.

[48] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2012.

[49] C. Minelli, A. De Grandi, C. X. Weichenberger, M. Gögele, M. Modenese, J. Attia, J. H. Barrett, M. Boehnke, G. Borsani, G. Casari, et al. Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genetic Epidemiology*, 37(2):205–213, 2013.

[50] P. J. Newcombe, D. V. Conti, and S. Richardson. JAM: a scalable Bayesian framework for joint analysis of marginal SNP effects. *Genetic epidemiology*, 40 (3):188–201, 2016.

[51] I. Ntzoufras. *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons, 2011.

[52] R. B. O'Hara, M. J. Sillanpää, et al. A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

[53] A. J. O'Malley, R. G. Frank, and S.-L. Normand. Estimating cost-offsets of new medications: Use of new antipsychotics and mental health costs for schizophrenia. *Statistics in Medicine*, 30(16):1971–1988, 2011.

[54] T. M. Palmer, D. A. Lawlor, R. M. Harbord, N. A. Sheehan, J. H. Tobias, N. J. Timpson, G. Davey Smith, and J. A. C. Sterne. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*, 21(3):223–242, Jun 2012. doi: 10.1177/0962280210394459.

[55] B. L. Pierce and S. Burgess. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184, 2013.

[56] B. L. Pierce, H. Ahsan, and T. J. Vanderweele. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology*, 40(3):740–752, Jun 2011. doi: 10.1093/ije/dyq151.

[57] D. E. Reich, S. B. Gabriel, and D. Altshuler. Quality and completeness of SNP databases. *Nature Genetics*, 33(4):457–458, 2003.

[58] Schizophrenia Psychiatric Genome-Wide Association Study Consortium et al. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, 43(10):969–976, 2011.

[59] C. Y. Shapland. *Many dependent instruments in Mendelian randomisation.* PhD thesis, University of Leicester, 2017.

[60] N. A. Sheehan, S. Meng, and V. Didelez. Mendelian randomisation: a tool for assessing causality in observational epidemiology. *Genetic Epidemiology*, pages 153–166, 2011.

[61] D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica: Journal of the Econometric Society*, pages 557–586, 1997.

[62] J. Stamler. Epidemiologic findings on body mass and blood pressure in adults. *Annals of Epidemiology*, 1(4):347–362, 1991.

[63] T. Strachan and A. Read. *Human Molecular Genetics 4.* Garland Science/Taylor & Francis Group, 2011.

[64] The International HapMap Consortium. The International HapMap project. *Nature*, 426(6968):789–796, 2003.

[65] D. C. Thomas and D. V. Conti. Commentary: the concept of 'Mendelian Randomization'. *International Journal of Epidemiology*, 33(1):21–25, Feb 2004. doi: 10.1093/ije/dyh048.

[66] J. R. Thompson, C. Minelli, K. R. Abrams, M. D. Tobin, and R. D. Riley. Meta-analysis of genetic studies using Mendelian randomization–a multivariate approach. *Statistics in Medicine*, 24(14):2241–2254, 2005.

[67] J. R. Thompson, M. Gögele, C. X. Weichenberger, M. Modenese, J. Attia, J. H. Barrett, et al. SNP prioritization using a Bayesian probability of association. *Genetic Epidemiology*, 37(2):214–221, 2013.

[68] N. J. Timpson, R. Harbord, G. Davey Smith, J. Zacho, A. Tybjaerg-Hansen, and B. G. Nordestgaard. Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype. *Hypertension*, 54(1):84–90, Jul 2009. doi: 10.1161/HYPERTENSIONAHA.109.130005.

[69] M. D. Tobin, M. Tomaszewski, P. S. Braund, C. Hajat, S. M. Raleigh, T. M. Palmer, M. Caulfield, P. R. Burton, and N. J. Samani. Common variants in genes underlying monogenic hypertension and hypotension and blood pressure in the general population. *Hypertension*, 51(6):1658–1664, Jun 2008. doi: 10.1161/HYPERTENSIONAHA.108.112664.

[70] F. Wang, N. J. Meyer, K. R. Walley, J. A. Russell, and R. Feng. Causal genetic inference using haplotypes as instrumental variables. *Genetic Epidemiology*, 40 (1):35–44, 2016.

[71] P. H. Whincup, S. J. Kaye, C. G. Owen, R. Huxley, D. G. Cook, S. Anazawa, et al. Birth weight and risk of type 2 diabetes: a systematic review. *JAMA*, 300(24):2886–2897, Dec 2008. doi: 10.1001/jama.2008.886.

[72] R. R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic Press, 2012.

[73] F. Windmeijer, H. Farbmacher, N. Davies, et al. On the use of the lasso for instrumental variables estimation with some invalid instruments. 2017.

[74] World Health Organization. *Obesity: preventing and managing the global epidemic*, volume 894. WHO technical report series, 1999.

[75] H. Yaghootkar, C. Lamina, R. A. Scott, Z. Dastani, M.-F. Hivert, L. L. Warren, et al. Mendelian randomisation studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. *Diabetes*, page DB_130128, 2013.

10 Supplementary tables and figures

10.1 Convergence and mixing

Table 8: Convergence Diagnostic by comparing 5 short chains with 1 long chain. The short chain had 50,000 iterations with 10,000 burn in. The long chain had 500,000 iterations and 250,000 burn in. The true β_{XY} is 0.2449. SE is Time-Series standard error. Prob. X is the probability of X included in the second regression. Total Visit. Prob. is the visited probability of the set of models chosen in the first regression; the set consist of the top 10 models from the longer chain.

Ins.	Chain	Mean $\hat{\beta}_{XY}$	SE	95% Credible Int.		Prob. X	Total Visit. Prob.
Number of Instruments (with variable MAF and positive confounding)							
10	1	0.0856	0.0154	-0.1613	0.5020	0.4448	0.6930
	2	0.0923	0.0155	-0.0900	0.4972	0.4351	0.6684
	3	0.1048	0.0188	-0.1454	0.5496	0.4651	0.6943
	4	0.1008	0.0152	-0.0679	0.5143	0.4494	0.6829
	5	0.1366	0.0180	-0.0779	0.5544	0.5313	0.6600
	Long	0.0995	0.0068	-0.1330	0.5259	0.4584	0.6786
30	1	0.2839	0.0220	0.0000	0.6645	0.7621	0.1061
	2	0.2766	0.0237	0.0000	0.6801	0.7274	0.1051
	3	0.2954	0.0183	0.0000	0.6603	0.8106	0.1015
	4	0.2536	0.0218	0.0000	0.6540	0.7174	0.1068
	5	0.2819	0.0181	0.0000	0.6529	0.7877	0.1056
	Long	0.2906	0.0079	0.0000	0.6597	0.7881	0.1034
60	1	0.2017	0.0222	-0.0142	0.6165	0.6219	0.0285
	2	0.2025	0.0243	-0.0614	0.6215	0.6291	0.0346
	3	0.2275	0.0198	-0.0162	0.6333	0.6947	0.0287
	4	0.2103	0.0277	-0.1478	0.6661	0.6597	0.0358
	5	0.1744	0.0245	-0.0912	0.6489	0.5621	0.0263
	Long	0.1801	0.0093	-0.0724	0.6077	0.5939	0.0315
90	1	0.2068	0.0226	-0.0157	0.6452	0.6356	0.0106
	2	0.2043	0.0202	-0.0107	0.5802	0.6602	0.0101
	3	0.2228	0.0207	0.0000	0.6164	0.6775	0.0121
	4	0.1659	0.0206	-0.0512	0.5933	0.5789	0.0137

Continued on next page

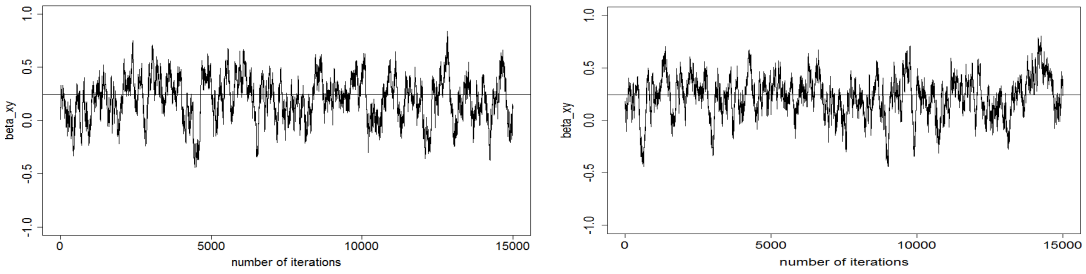
Table 8 – *Continued from previous page*

Ins.	Chain	Mean $\hat{\beta}_{XY}$	SE	95% Credible Int.		Prob. X	Total Visit. Prob.
	5	0.2154	0.0252	-0.0306	0.6510	0.6496	0.0128
	Long	0.1962	0.0096	-0.0377	0.6236	0.6210	0.0137
MAF (with 10 instruments and positive confounding)							
Low	1	0.4145	0.0161	0.0000	0.7490	0.9104	0.5053
	2	0.4226	0.0157	0.0000	0.7543	0.9148	0.4861
	3	0.3855	0.0178	0.0000	0.7225	0.8771	0.4956
	4	0.4338	0.0140	0.0000	0.7367	0.9240	0.5132
	5	0.4167	0.0148	0.0000	0.7340	0.9192	0.5056
	Long	0.4272	0.0065	0.0000	0.7473	0.9252	0.5069
Variable	1	0.0856	0.0154	-0.1613	0.5020	0.4448	0.6930
	2	0.0923	0.0155	-0.0900	0.4972	0.4351	0.6684
	3	0.1048	0.0188	-0.1454	0.5496	0.4651	0.6943
	4	0.1008	0.0152	-0.0679	0.5143	0.4494	0.6829
	5	0.1366	0.0180	-0.0779	0.5544	0.5313	0.6600
	Long	0.0995	0.0068	-0.1330	0.5259	0.4584	0.6786
Confounding effect (with 10 instruments and variable MAF)							
Positive	1	0.0856	0.0154	-0.1613	0.5020	0.4448	0.6930
	2	0.0923	0.0155	-0.0900	0.4972	0.4351	0.6684
	3	0.1048	0.0188	-0.1454	0.5496	0.4651	0.6943
	4	0.1008	0.0152	-0.0679	0.5143	0.4494	0.6829
	5	0.1366	0.0180	-0.0779	0.5544	0.5313	0.6600
	Long	0.0995	0.0068	-0.1330	0.5259	0.4584	0.6786
Negative	1	0.4312	0.0297	0.0000	1.0558	0.8951	0.7213
	2	0.4064	0.0270	0.0000	0.9535	0.8576	0.7181
	3	0.4292	0.0240	0.0000	0.9532	0.8977	0.7400
	4	0.3777	0.0349	0.0000	0.9754	0.7804	0.7271
	5	0.4680	0.0304	0.0000	1.0914	0.9116	0.7287
	Long	0.4379	0.0115	0.0000	1.0038	0.8842	0.7249
Strong	1	0.0849	0.0235	-0.2470	0.4869	0.4556	0.7671
	2	0.0530	0.0264	-0.3916	0.4592	0.4309	0.7782
	3	0.0727	0.0171	-0.1352	0.4800	0.4050	0.7859

Continued on next page

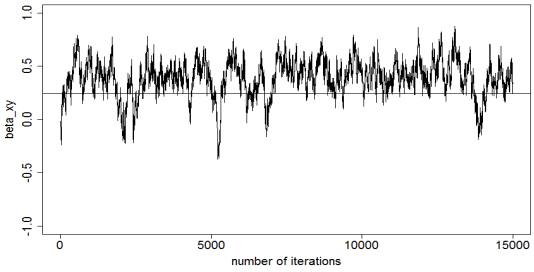
Table 8 – *Continued from previous page*

Ins.	Chain	Mean $\hat{\beta}_{XY}$	SE	95% Credible Int.		Prob. X	Total Visit. Prob.
	4	0.0923	0.0277	-0.3282	0.5217	0.4802	0.7617
	5	0.0908	0.0242	-0.3049	0.4971	0.4953	0.7667
	Long	0.0797	0.0079	-0.1404	0.4704	0.4312	0.7774

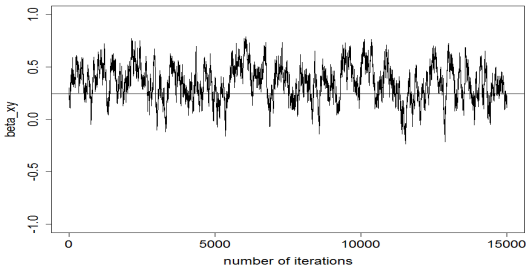


(a) 10 Instruments & Short Chain

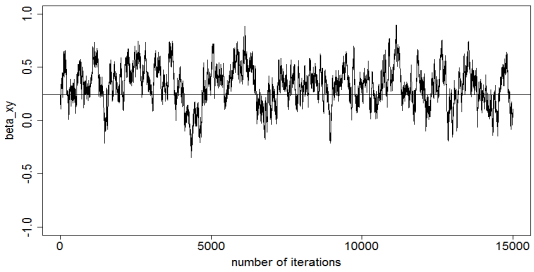
(b) 10 Instruments & Long Chain



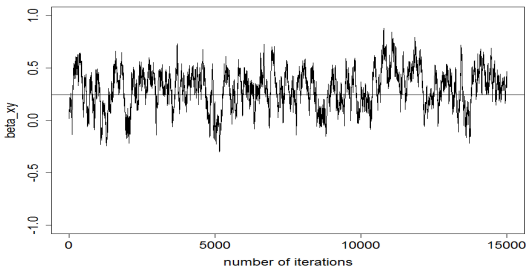
(c) 30 Instruments & Short Chain



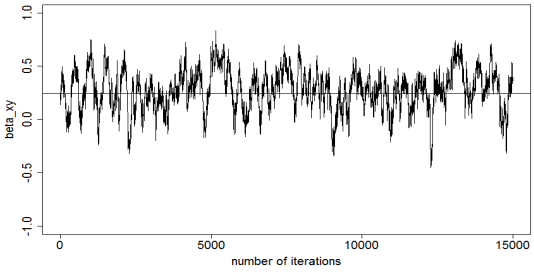
(d) 30 Instruments & Long Chain



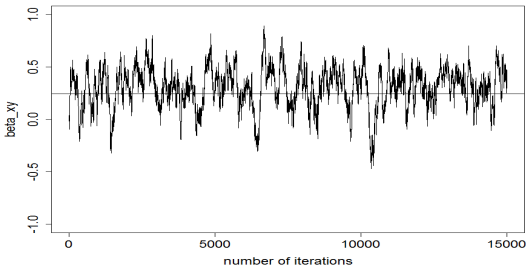
(e) 60 Instruments & Short Chain



(f) 60 Instruments & Long Chain

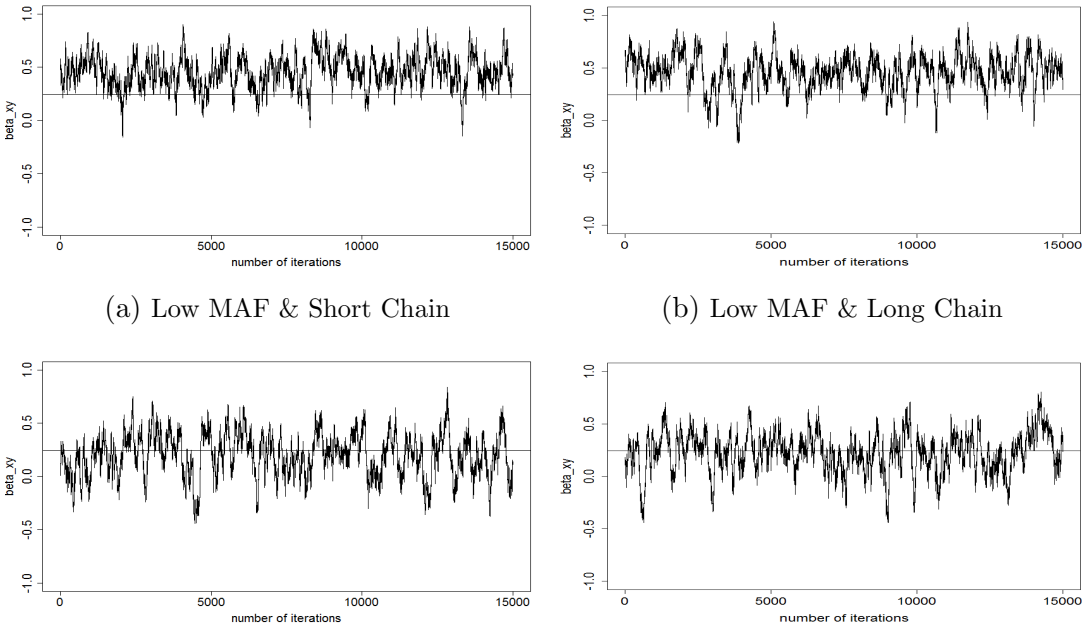


(g) 90 Instruments & Short Chain



(h) 90 Instruments & Long Chain

Figure 4: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10,30,60 and 90 instruments. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



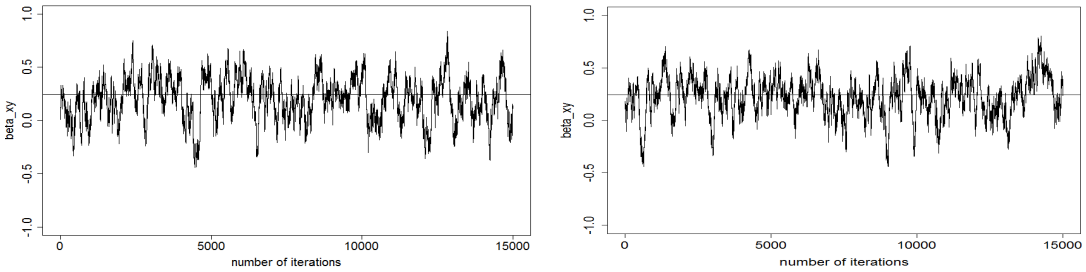
(a) Low MAF & Short Chain

(b) Low MAF & Long Chain

(c) 10 Instruments & Short Chain

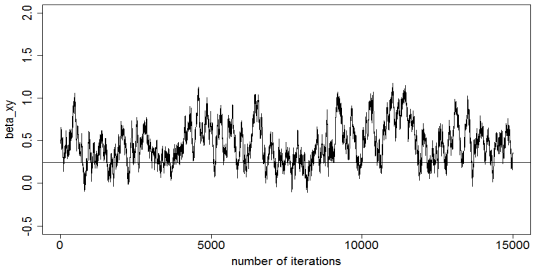
(d) 10 Instruments & Long Chain

Figure 5: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with different MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

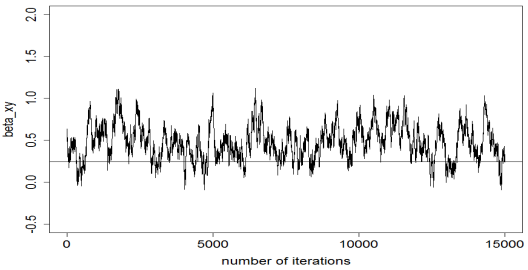


(a) Positive Confounding & Short Chain

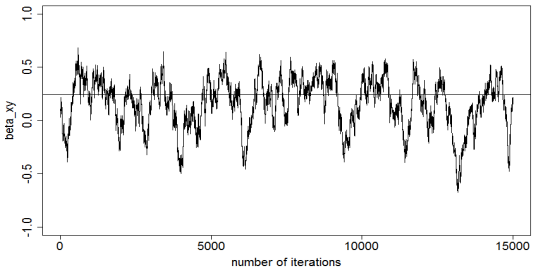
(b) Positive Confounding & Long Chain



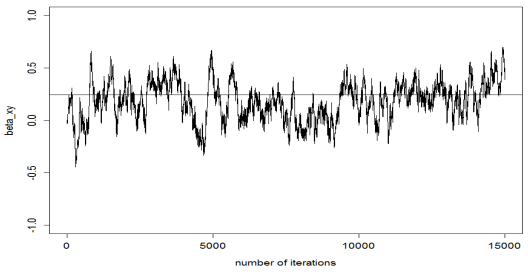
(c) Negative Confounding & Short Chain



(d) Negative Confounding & Long Chain



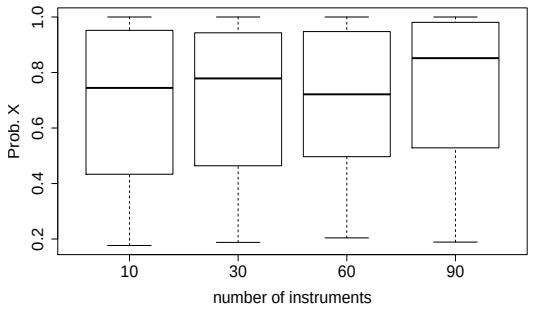
(e) Strong Confounding & Short Chain



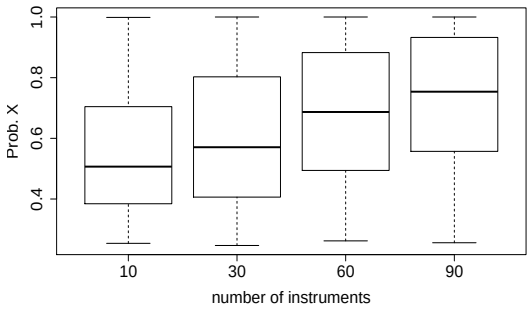
(f) Strong Confounding & Long Chain

Figure 6: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with different confounding effect with short and long chain. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

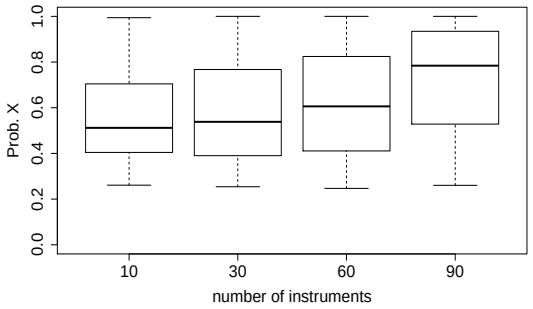
10.2 A comparison to the classical estimators



(a) Common MAF

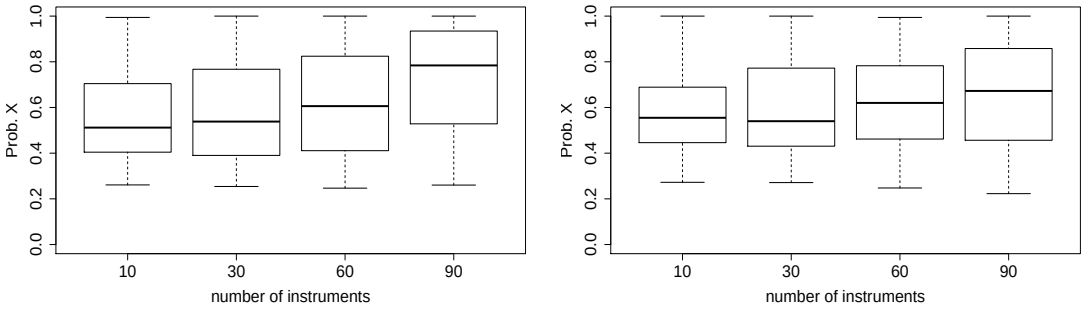


(b) Low MAF



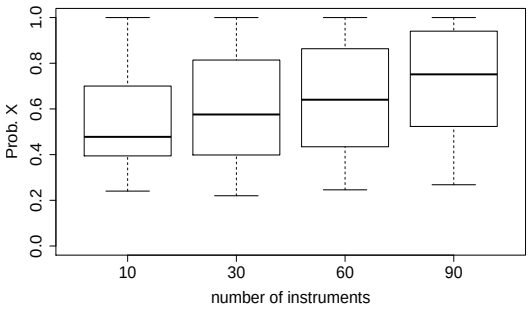
(c) Variable MAF

Figure 7: Probability of X for common, variable and low MAF with different number of instruments from *ivbma*. Probability of X, is the probability of X included in the second regression of IVBMA.

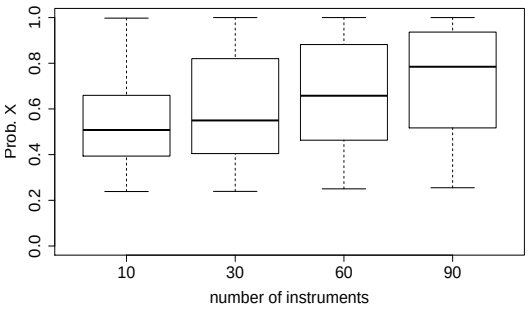


(a) Pattern I

(b) Pattern II



(c) Pattern III



(d) Pattern IV

Figure 8: Probability of X for all four patterns with different number of instruments from *ivbma*. Probability of X, is the probability of X included in the second regression of IVBMA.